**Regression – Assignment**

Predict the insurance charges based on the several parameters.

1. Identify the problem statement
2. Stage 1 – Domain Selection – Numerical Values in the Dataset

Machine Learning

1. Stage 2 – Learning Selection – Input and Output are well defined

Supervised Learning

1. Stage 3 – Regression or Classification – Continuous numerical value

Regression

1. Basic info about the dataset

The dataset which is given by the clients contains 6 columns, and 1338 rows. The first 5 columns which are ‘age’, ‘sex’, ‘BMI’, ‘children’ and ‘smoker’ are considered as features. The last column ‘charges’ is considered as target. The inputs of the output are well defined in the database.

1. Data Preprocessing

In the data set the column ‘sex’ and ‘smoker’ are categorical columns. We cannot process the categorical data in python, we need to convert it into numerical data.

Multiple Linear Algorithm

R score value - 0.7100

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Support Vector Machine Algorithm | | | |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Serial Number** | **C Value** | **Linear** | **RBF** | **Poly** | **Sigmoid** |
| **(R\_Value)** | **(R\_Value)** | **(R\_Value)** | **(R\_Value)** |
| 1 | 100 | -0.6250 | -14.6720 | -12.4292 | -319.3378 |
| 2 | 500 | -0.005 | -11.9412 | -11.6927 | -9.5046 |
| 3 | 700 | 0.00829 | -11.7580 | -11.2107 | -3.9962 |
| 4 | 1000 | 0.0406 | -11.6079 | -10.7178 | -1.9844 |
| 5 | 2000 | 0.3405 | -11.2402 | -9.1897 | -0.6638 |
| 6 | 3000 | 0.6615 | -10.9052 | -7.832 | -0.3690 |

SVM R score value - 0.661500

Decision Tree Algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Serial Number** | **CRITERION** | **MAX FEATURES** | **SPLITTER** | **R SCORE** |
| 1 | Squared\_error (mse) | Auto | Best | 0.7171 |
| 2 | Squared\_error | Auto | Random | 0.7210 |
| 3 | Squared\_error | Sqrt | Best | 0.6420 |
| 4 | Squared\_error | Sqrt | Random | 0.6731 |
| 5 | Squared\_error | Log2 | Best | 0.7344 |
| 6 | Squared\_error | Log2 | Random | 0.7502 |
| 7 | Absolute\_error (mae) | Auto | Best | 0.7219 |
| 8 | Absolute\_error | Auto | Random | 0.7525 |
| 9 | Absolute\_error | Sqrt | Best | 0.7599 |
| 10 | Absolute\_error | Sqrt | Random | 0.6433 |
| 11 | Absolute\_error | Log2 | Best | 0.7064 |
| 12 | Absolute\_error | Log2 | Random | 0.7203 |
| 13 | Friedman\_mse | Auto | Best | 0.7218 |
| 14 | Friedman\_mse | Auto | Random | 0.7678 |
| 15 | Friedman\_mse | Sqrt | Best | 0.7357 |
| 16 | Friedman\_mse | Sqrt | Random | 0.7489 |
| 17 | Friedman\_mse | Log2 | Best | 0.7415 |
| 18 | Friedman\_mse | Log2 | Random | 0.6535 |

Decision Tree R Score value - 0.7678

**Random Forest Algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Serial Number | CRITERION | MAX FEATURES | N\_ESTIMATORS | R SCORE |
| 1 | Squared\_error | Auto | 100 | 0.8534 |
| 2 | Squared\_error | Auto | 50 | 0.8519 |
| 3 | Squared\_error | Sqrt | 100 | 0.8534 |
| 4 | Squared\_error | Sqrt | 50 | 0.8572 |
| 5 | Squared\_error | Log2 | 100 | 0.8687 |
| 6 | Squared\_error | Log2 | 50 | 0.8653 |
| 7 | Absolute\_error | Auto | 100 | 0.8527 |
| 8 | Absolute\_error | Auto | 50 | 0.8513 |
| 9 | Absolute\_error | Sqrt | 100 | 0.8705 |
| 10 | Absolute\_error | Sqrt | 50 | 0.8693 |
| 11 | Absolute\_error | Log2 | 100 | 0.8690 |
| 12 | Absolute\_error | Log2 | 50 | 0.8689 |

Random forest R Score value - 0.8705

**Final Model**

I choose the **Random Forest Alogrithm** as the best model with passing parameters criterion= 'absolute\_error', max\_features= 'sqrt' and n\_estimators= 100. It’s R\_score value is much higher than the other algorithms and close to value 1.